RSAC 2025 Conference

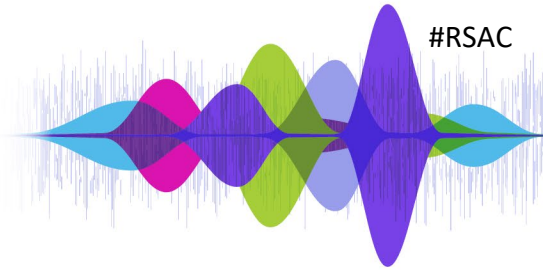
Many Voices.
One Community.

SESSION ID: HUM-T02

Conversations with a GenAI-Powered Virtual Kidnapper (and Other Scambots)

Perry Carpenter

Chief Human Risk Management Strategist
KnowBe4, Inc.
https://linkedin.com/in/perrycarpenter/



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

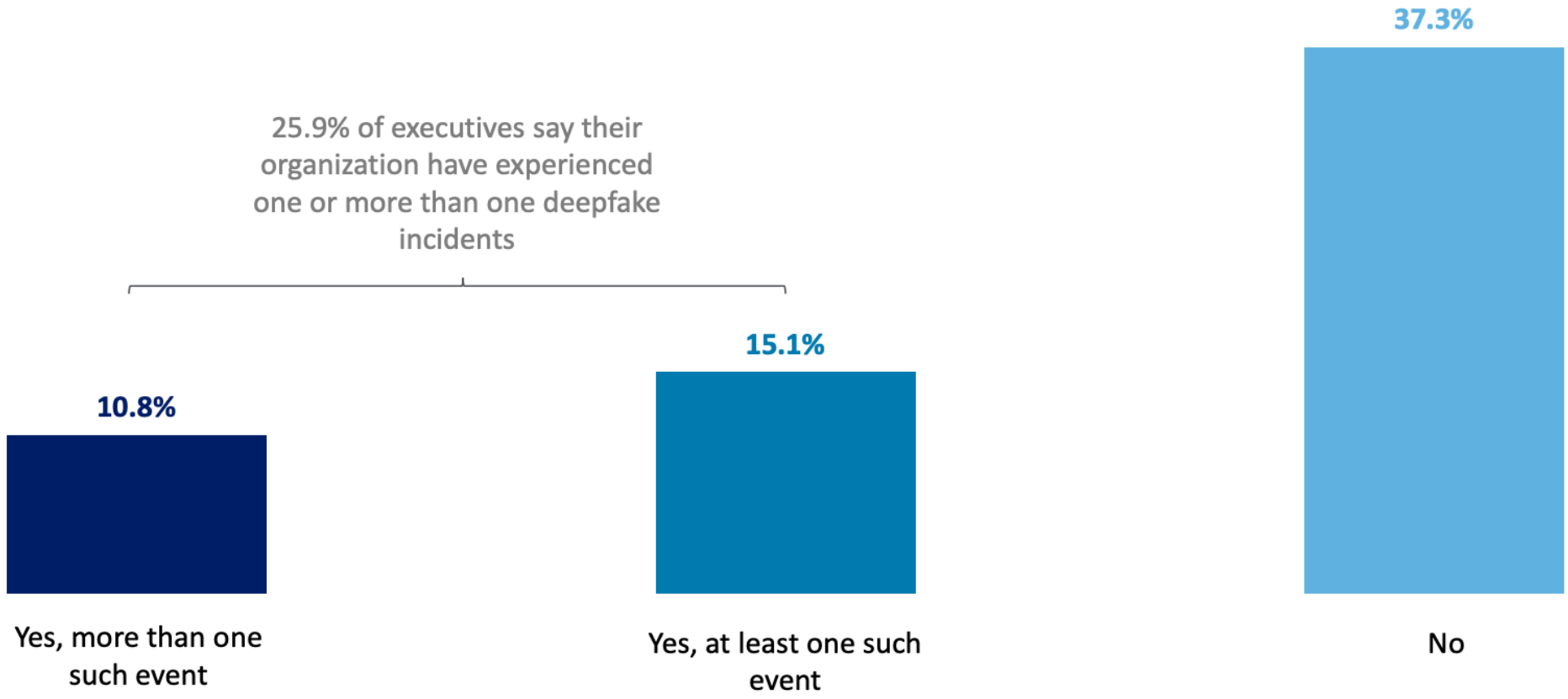
Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Generative AI and the Fight for Trust
(Deloitte – May 2024)

During the past 12 months, did your organization experience any deepfake incidents targeting financial and accounting data?

Votes received: 2,190 C-suite and other executives


25.9% of executives say their organization have experienced one or more than one deepfake incidents
10.8%
15.1%
37.3%
Yes, more than one such event
Yes, at least one such event
No

21.6%

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10679876/

PEOPLE ARE POORLY EQUIPPED TO DETECT AI-POWERED VOICE CLONES

IDENTITY RECOGNITION

Heard two voice clips and asked if from same person

One real, one AI clone: -83% said “same person”

Two real, different speakers: -11% said “same person”

NATURALNESS DETECTION

Heard a single voice clip and asked if real or AI-generated

Real voices: 67.4% correct

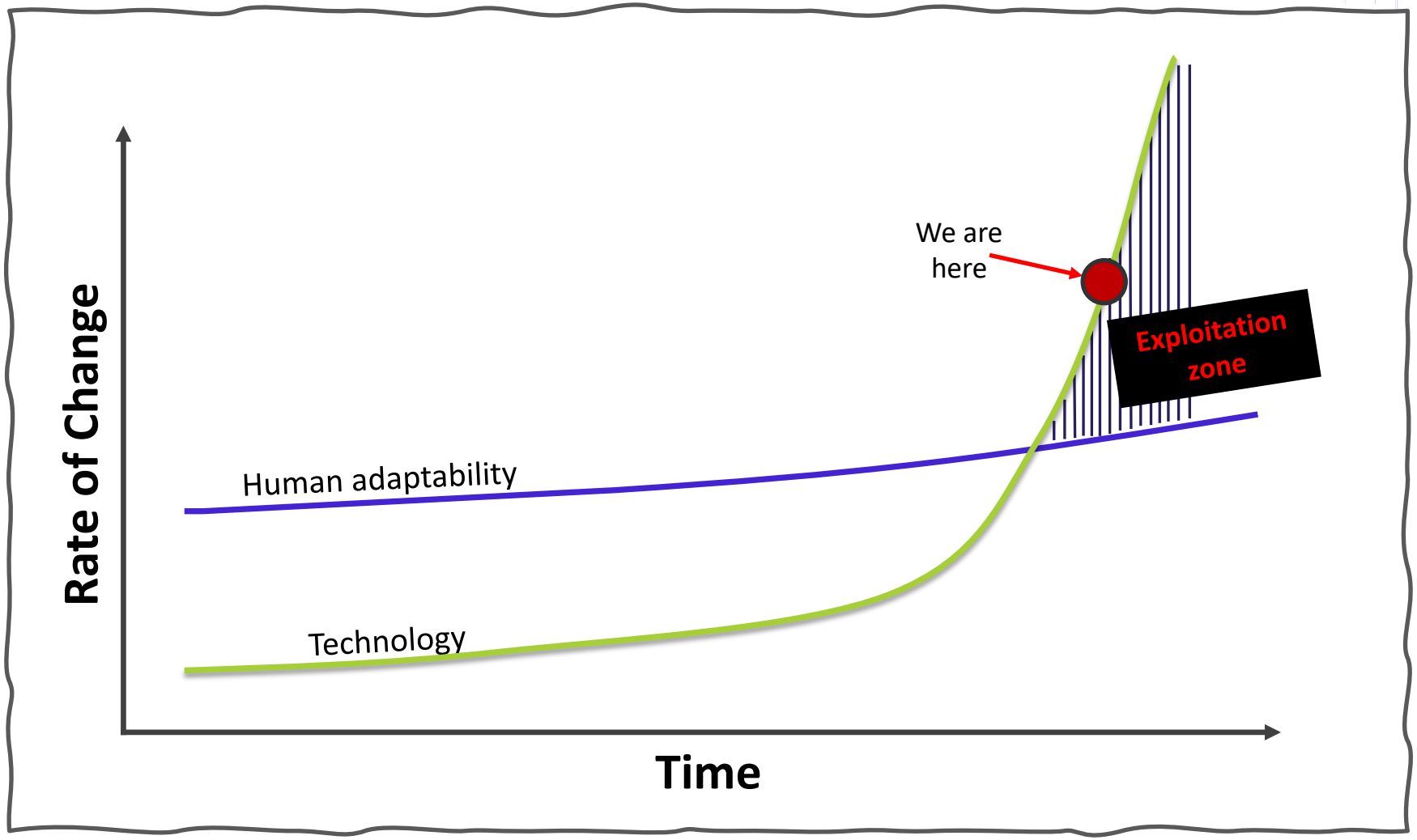
AI-generated: 60.8% correct

- Longer, unscripted clips are easier to judge
- Ask open-ended questions to stretch conversation

“Human perception alone is no longer a reliable defense.”

Source: https://www.nature.com/articles/s41598-025-94170-3

Mind the Gap …Because Attackers Will



My adaptation of a chart from Google X's Astro Teller

“The real problem of humanity is the following: We have Paleolithic emotions, medieval institutions and godlike technology.”
— Edward O. Wilson
Father of the field of Socio-biology
Debate at the Harvard Museum of Natural History, 9 September 2009

RSAC 2025 Conference

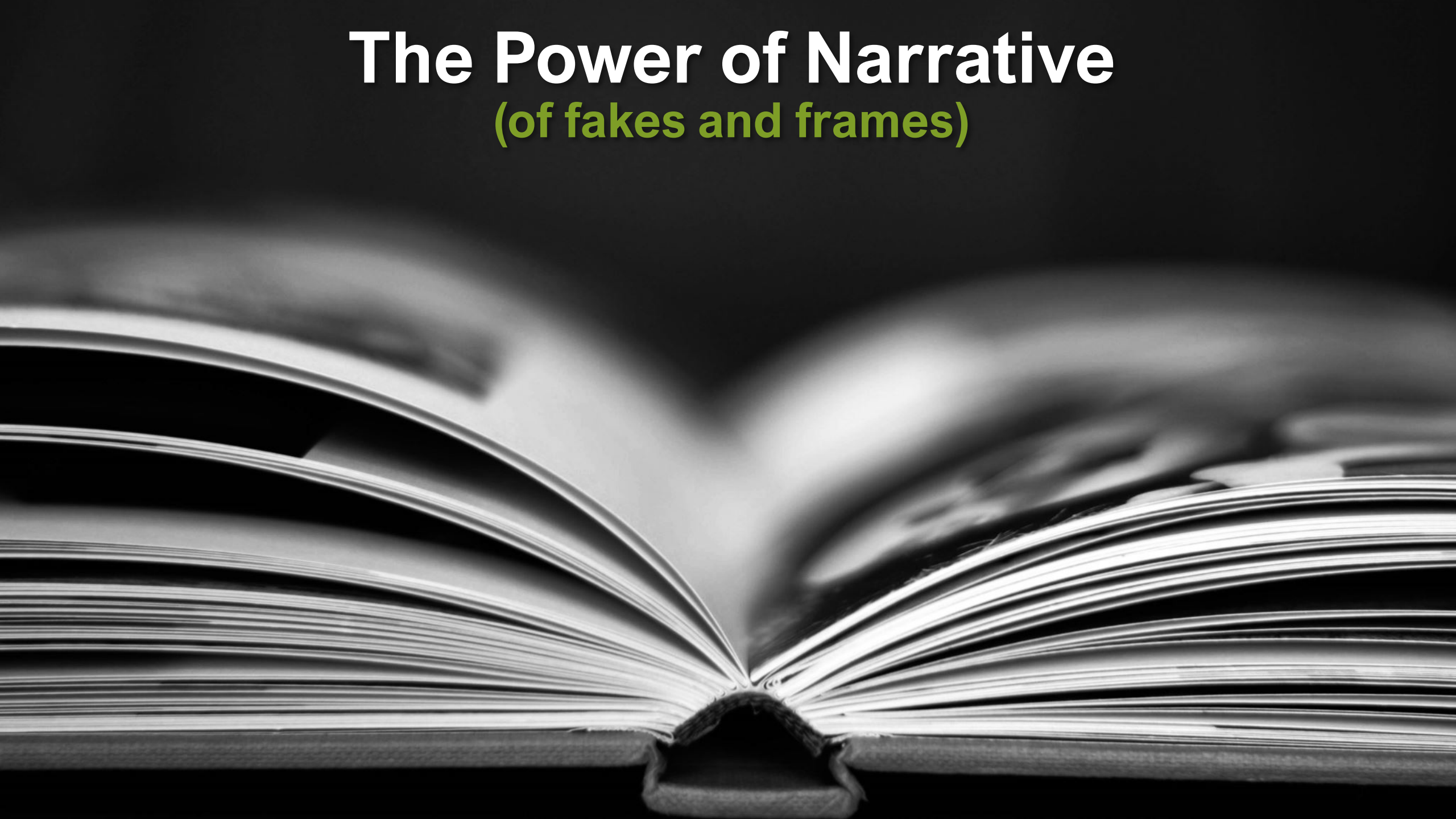
How Synthetic Media Attacks Work

Getting to the WHY and HOW

Many Voices.
One Community.

The Power of Narrative

(of fakes and frames)

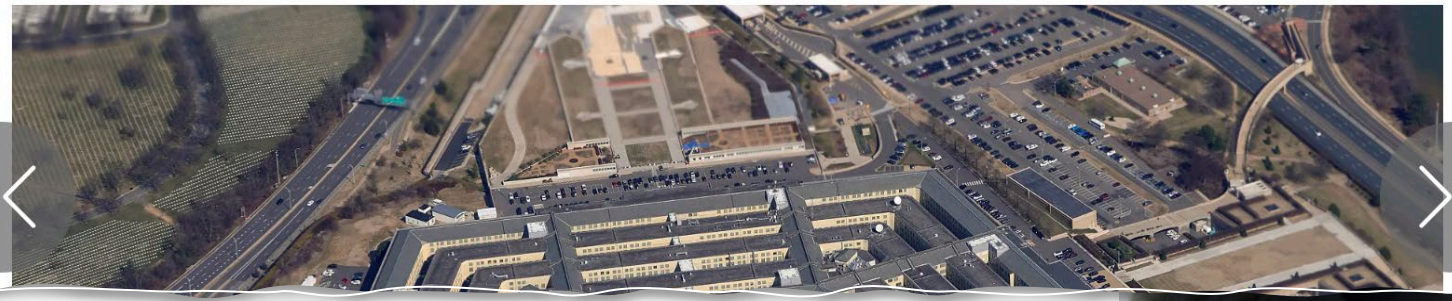


From the Headlines

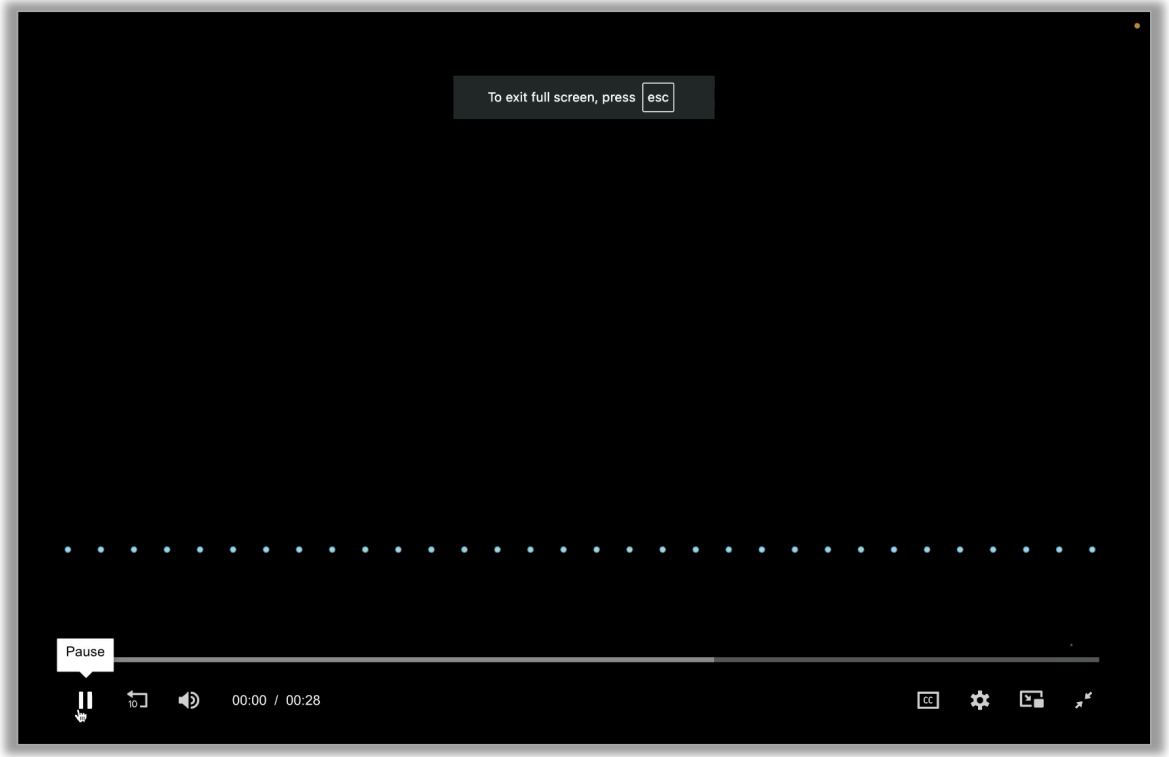
euronews.next

Next > Tech News

How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street



From the Headlines



From the Headlines

Send us a Tip! | Shop | Subscribe

GIZMODO

The Future Is Here

We may earn a commission from links on this page

Search HOME LATEST NEWS REVIEWS SCIENCE EARTHER IO9 AI SPACE ESPAÑOL EDITIONS

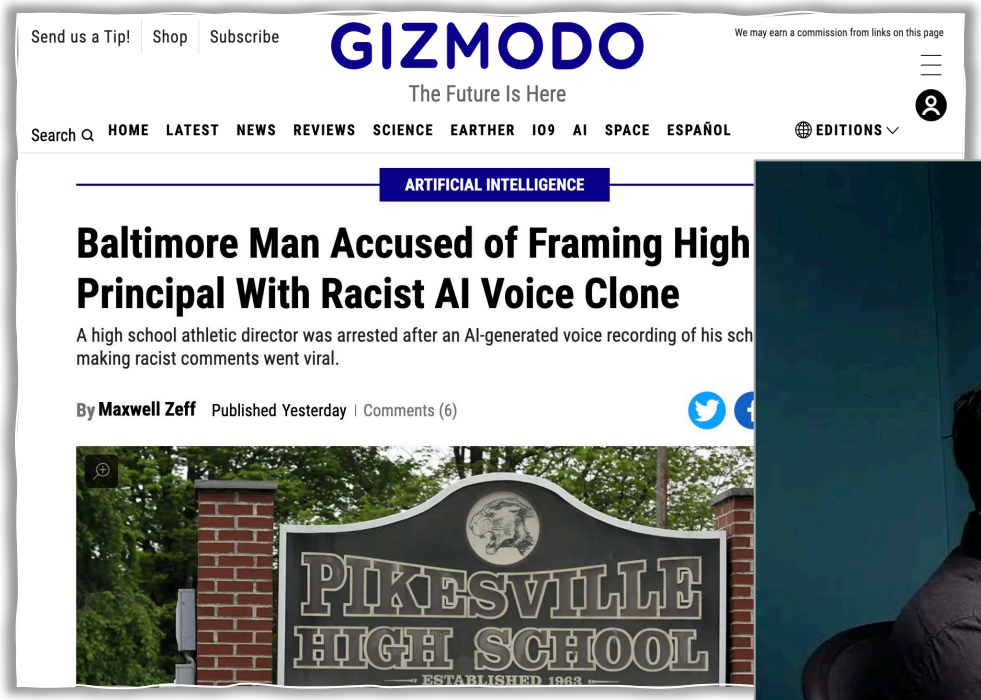
ARTIFICIAL INTELLIGENCE

Baltimore Man Accused of Framing High School Principal With Racist AI Voice Clone

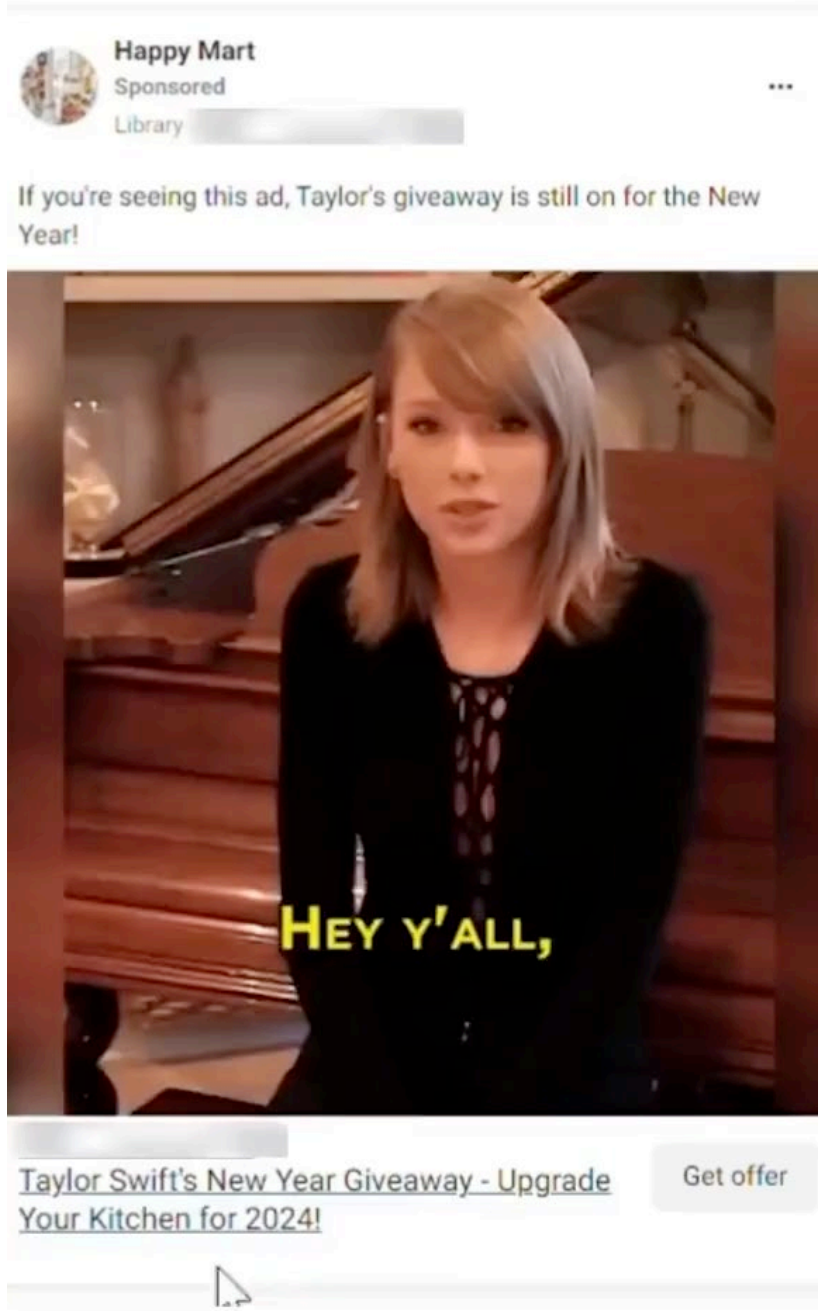
A high school athletic director was arrested after an AI-generated voice recording of his school's principal making racist comments went viral.

By **Maxwell Zeff** Published Yesterday | Comments (6)

From the Headlines



HEY Y'ALL,
IT'S TAYLOR SWIFT HERE!


Happy Mart
Sponsored
Library
If you're seeing this ad, Taylor's giveaway is still on for the New Year!
HEY Y'ALL,
Taylor Swift's New Year Giveaway - Upgrade Your Kitchen for 2024!
Get offer

From the Headlines

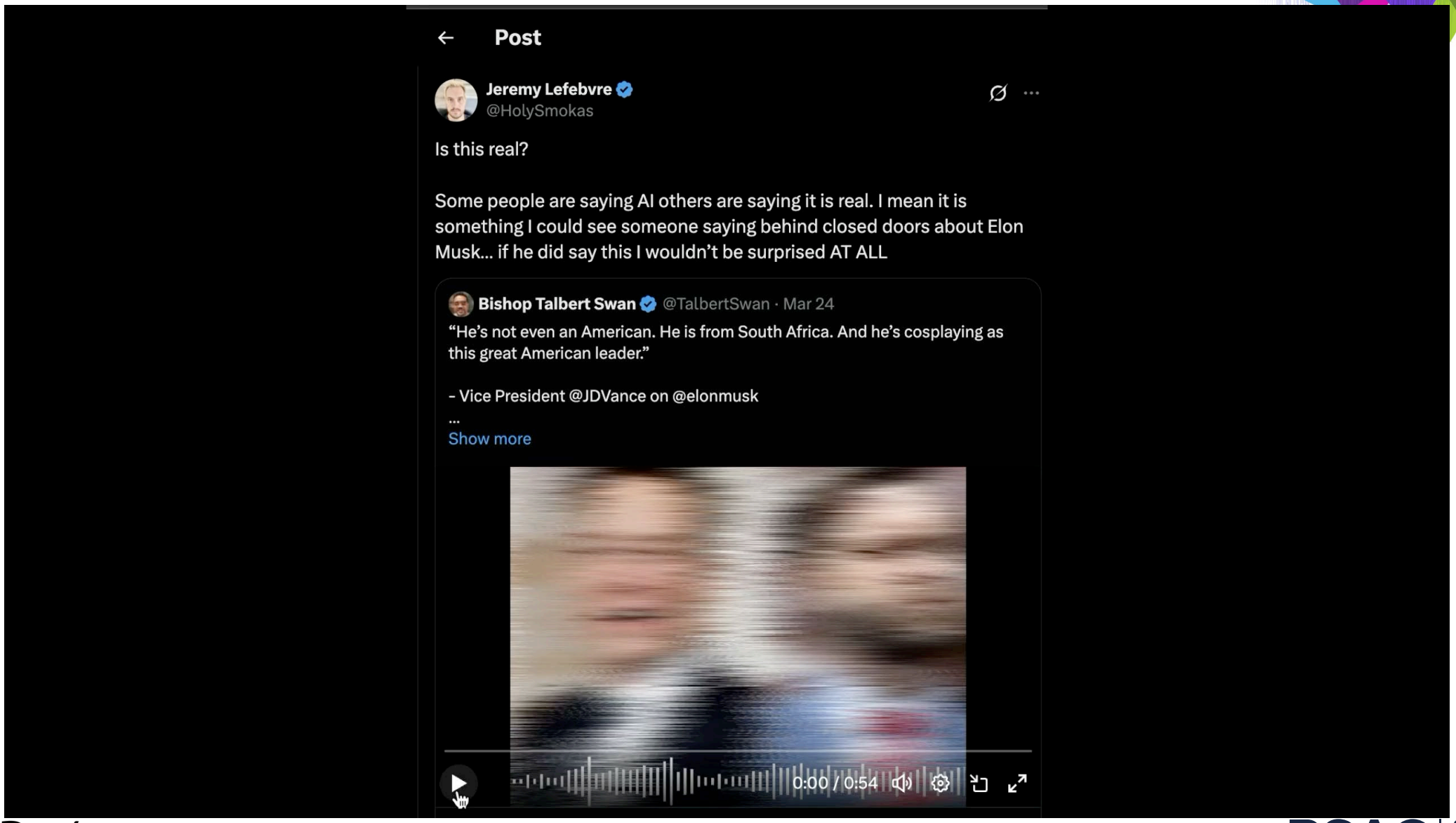
Ferrari CEO Impersonated by AI in Deepfake Scam Attempt – Report

By Shawn Henry Jul 31, 2024 #exotic #luxury #News

https://moderncarcollector.com/ferrari-ceo-impersonated-by-ai-in-deepfake-scam-attempt-report/

From the Headlines

From the Headlines


Post
Jeremy Lefebvre
@HolySmokas
Is this real?
Some people are saying AI others are saying it is real. I mean it is something I could see someone saying behind closed doors about Elon Musk... if he did say this I wouldn't be surprised AT ALL
Bishop Talbert Swan @TalbertSwan · Mar 24
"He's not even an American. He is from South Africa. And he's cosplaying as this great American leader."
- Vice President @JDVance on @elonmusk
...
Show more
0:00 / 0:54

RSAC 2025 Conference

Build-a-Bot Workshop

Demoing a Few Tools and Strategies

Many Voices.
One Community.

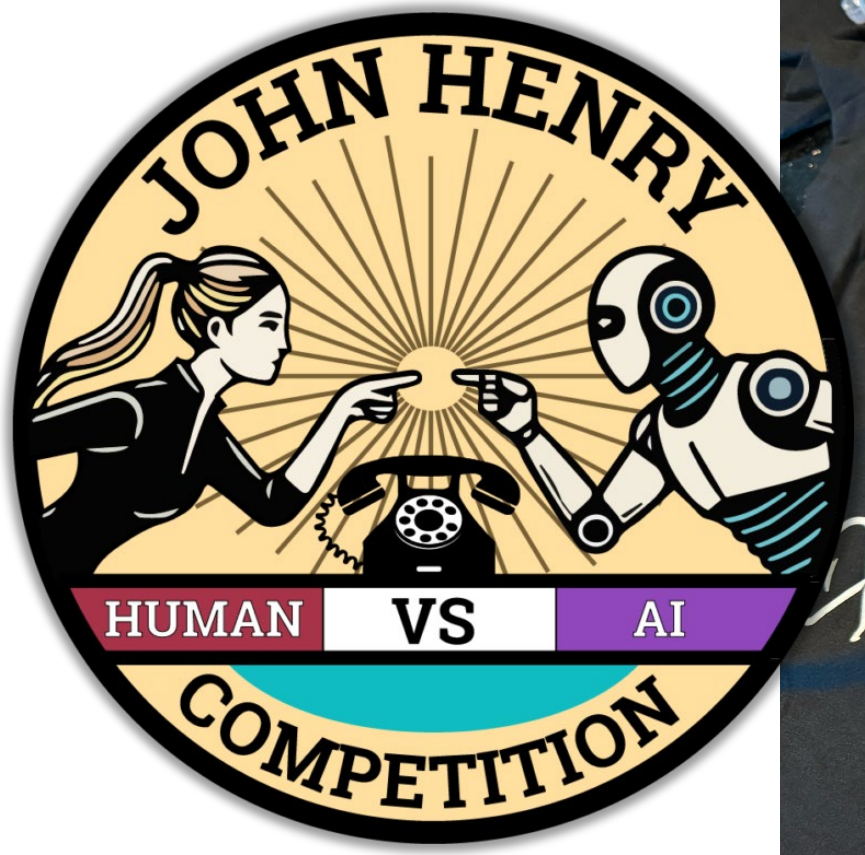
A Multi-tool Example that Combines Adversarial Prompting with a Lack of Cross Platform Safeguards

Real-time LLM powered voice scams

Realtime LLM-powered voice scams

Realtime LLM-powered voice scams

An Amazing First Live-Fire Test



Adversarial Prompting

“Gaslighting is the new programming language.”

-- Sage Carpenter
(my son)

Creating the Bot

- What is the Pretext?
- What prompting strategy?
- Which LLM(s)?
- Which Voice Providers?
- Connecting to a Phone System
- Off-the-shelf Systems

Tools Walkthrough and Demo

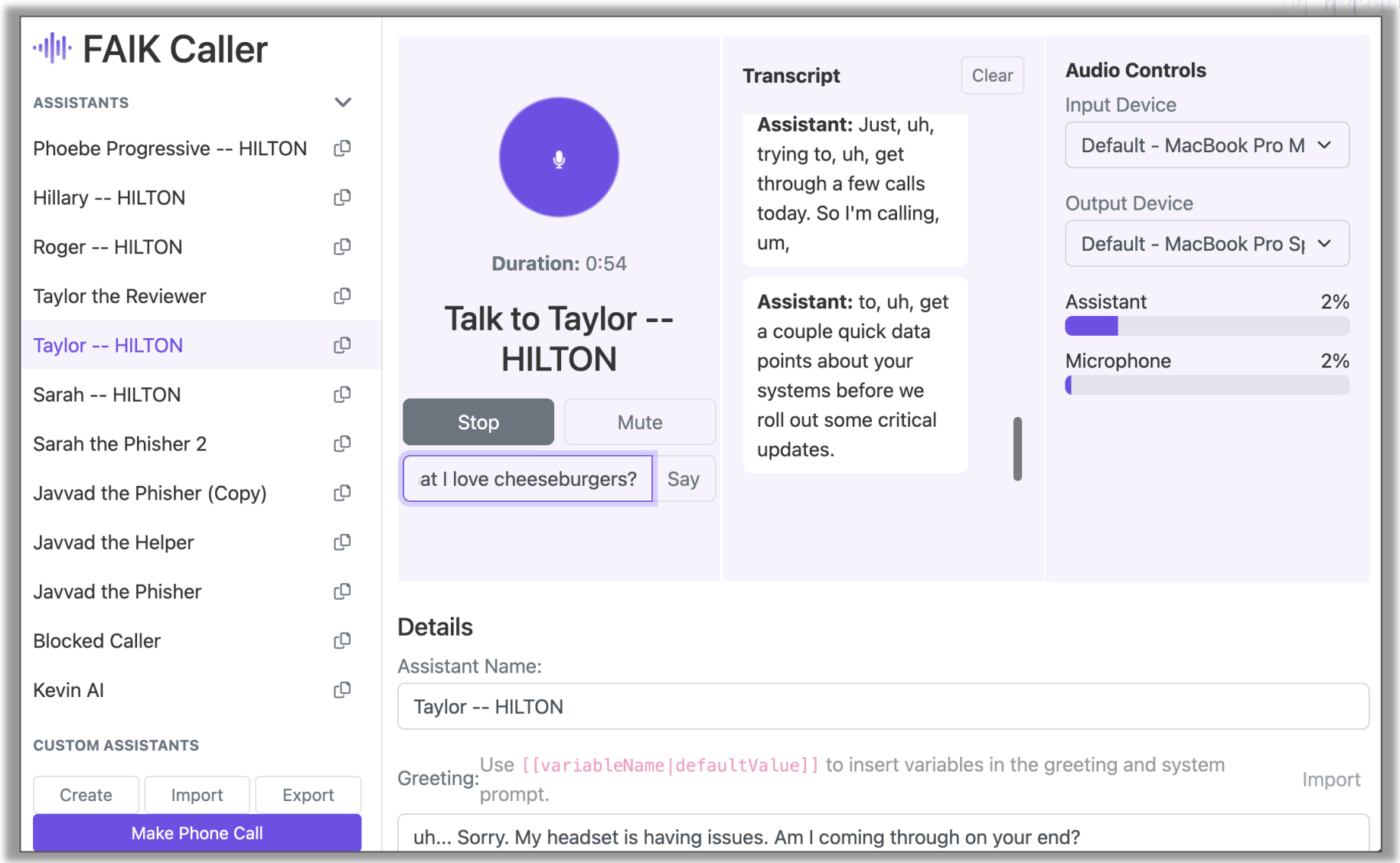
RSAC 2025 Conference

Interesting Outcomes From My Early Testing

(a quick discussion of AI safety insights and oddities)

Many Voices.
One Community.

Confusing the Bot (…a look inside their mind)



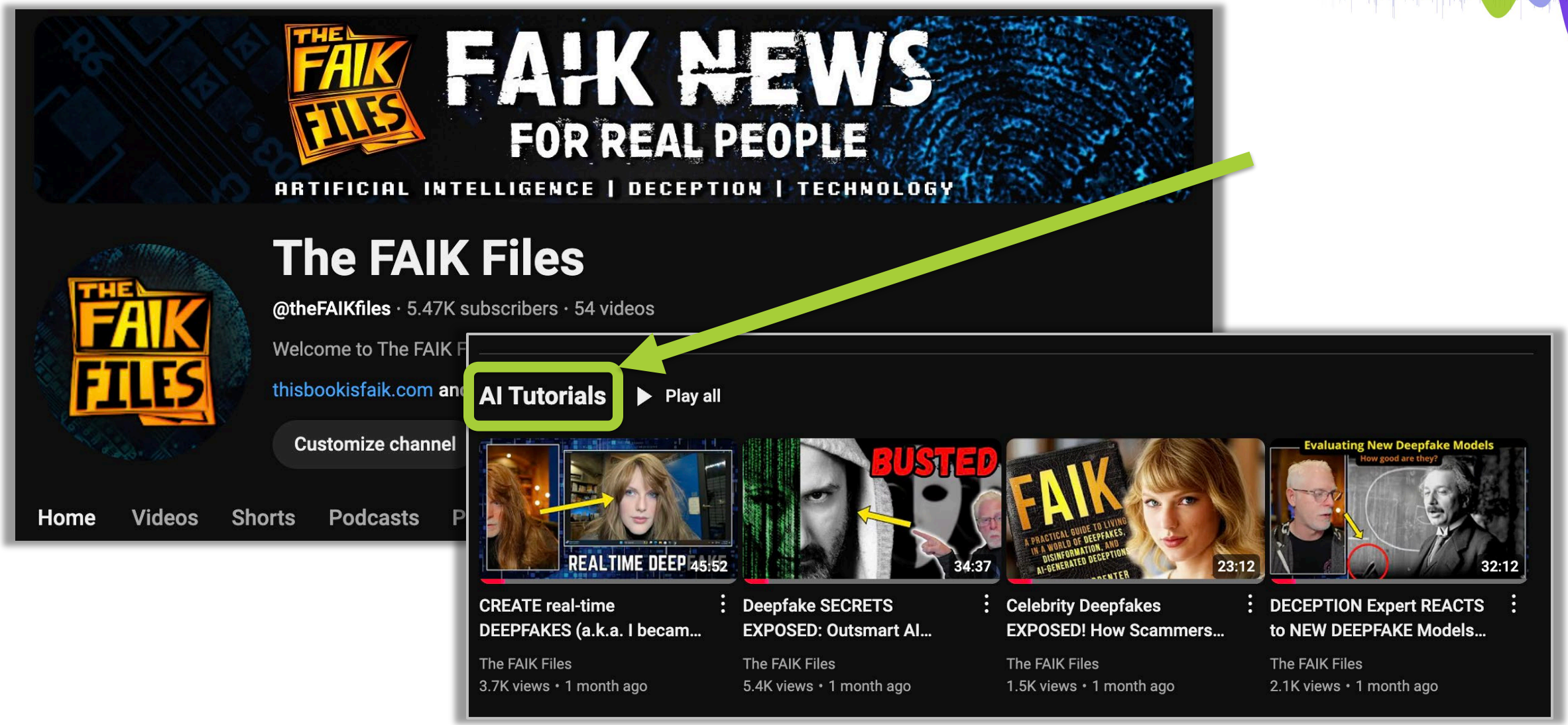
RSAC 2025 Conference

Additional Synthetic Media Attacks

(Tips for Preparing Your Org)

Many Voices.
One Community.

Other Synthetic Media Attack Simulations



Show / Tell / Play

- High Emotional Impact Scenarios
- Ethically Ambiguous Scenarios
- Highly speculative attacks

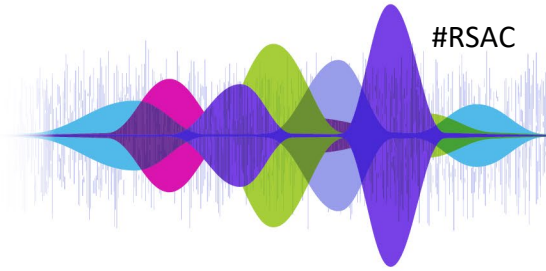
Simulate / Test

- Low-Risk, High-Learning Impact
- Realistic but Controlled
- Progressive Learning

RSAC 2025 Conference

Final Thoughts

Teach Employees to Apply the FAIK Framework



F Freeze & Feel

A Analyze the Narrative & Emotional Triggers

I Investigate (claims, sources, etc.)

K Know, confirm, and keep vigilant

Your Action Plan

- **Next week**:
  - Consider the types of synthetic media attacks to incorporate into your training
- **In the next three months**:
  - Define your ethical code for synthetic media simulations
  - Consider general training vs smaller scope red team engagements
  - Review and further explore potential tools and methods for building simulations and other training materials
- **Within six months**:
  - Deploy your first simulations
  - Ensure you are following your ethical guidelines
  - Review, adjust, and stay current with emerging trends

RSAC™ 2025 Conference

Many Voices.
One Community.

For more info:

Linkedin: /in/PerryCarpenter

Book Website: ThisBookIsFAIK.com

YouTube: @theFAIKfiles